PERFORMANCE EVALUATION OF INACT - INDECT ADVANCED IMAGE CATALOGUING TOOL

Libor MICHALEK[1], *Michal GREGA*[2], *Damian BRYK*[2], *Bartlomiej GRABOWSKI*[2]

[1]Department of Telecommunications, Faculty of Electrical Engeneering and Computer Science, VSB-Technical University Ostrava, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
[2]Department of Telecommunications, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. Mickiewicza 30, 30 059 Krakow, Poland

libor.michalek@vsb.cz, grega@kt.agh.edu.pl, dmn.bryk@gmail.com, grabbartek@gmail.com

Abstract. *In this article, we describe the performance evaluation of INACT tool which is developed for cataloguing of high-level and low-level metadata of the evidence material. INACT tool can be used by police forces in the cases of prosecution of such crimes as as possession and distribution of child pornography (CP). In live forensic cases, the time to first hit (time when the first image containing e.g. CP is found) is important, as then further legal actions are justified (such as arrest of the suspect and his hardware). The performance evaluation of first hit was performed on real data with the cooperation of Czech Police, Department of Internet Crime.*

Keywords

Forenscis, child pornography, INACT, INDECT, performance evaluation, Police Tools.

1. Introduction

Child pornography (CP) is a term that broadly describes all multimedia and writings that depict sexual activities involving a child. As at 2008, 94 of 187 Interpol member states refer to CP as a crime in their code of laws. Of these, possession of such content is punishable in 58 countries [1]. Since this crime is regarded as being extremely harmful, its prosecution is of the highest priority for police forces and law enforcement organizations around the world. The easiest way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

Possession of CP images is considered a crime in most countries. This law applies not only to regular citizens, but also to police units. Local regulations force police officers to destroy all evidence after the investigation is completed. This significantly complicates the gathering and presenting of evidence in police investigations. It also makes it impossible for police officers to make connections between individual cases in order to track the distribution paths of such content. This presents a requirement for computer software to overcome this problem.

The result of the research and development is the INACT (INDECT Advanced Image Catalogue Tool) software, see [2], [3]. This article follows up [2] by practical experiments and performance evaluation. The practical experiments were performed with the cooperation of Czech Police, Department of Internet Crime.

2. Concept of the Application

In order to solve the problem outlined in the introduction, a set of two complimentary applications was developed. The first application, the INACT INDEXER, is designed to be used at police stations. The police have at their disposal sets of images containing child pornography from various sources, including ongoing investigations and international channels of cooperation. Such sets are input into the INACT INDEXER. The INACT INDEXER processes the images and creates a catalogue containing information about the images (such as a description of the case in which the images were acquired) and a set of descriptors for each of the catalogued images. The descriptor set consists of MD5 hashes and MPEG-7 descriptors. The process of calculating hashes and descriptors is a one-way process, which means that the images cannot be recreated either from the hashes or from the descriptors. This allows the police to dispose of the images (as required by law) whilst retaining the information about the images themselves. The result of the INACT INDEXER analysis is a database which can be utilized by the INACT INSEARCHER application.

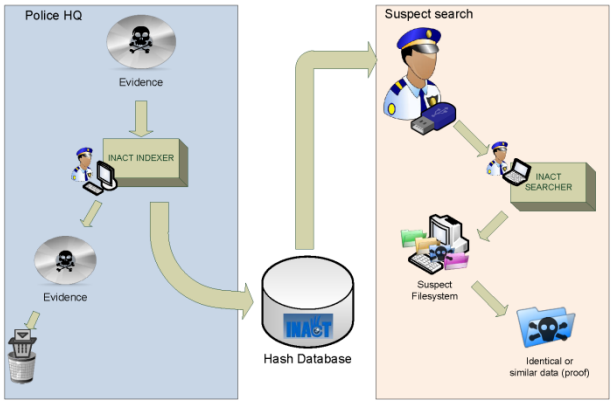

Fig. 1: INACT concept.

The database is centralized, which allows multiple instances of the INACT INDEXER to be run at the same time. This allows for country-wide deployment of the application (e.g. to all regional police forces) while retaining a coherent database of the hashes. The concept of the two INACT applications is presented in Fig. 1. For more information about the INTACT concept including also system requirements see [2].

3. INACT INDEXER

According to the current legal regulations regarding police investigations in numerous Interpol countries, police units cannot store evidence containing prohibited content. Any data carrier with illegal images or videos must be destroyed after it is analyzed. Even if the police were permitted to gather this content, it is physically difficult to store large amounts of multimedia content. One of the INACT project goals is to create a tool which can automatically analyze illegal files and store metadata extracted during this procedure in a database. The database should consist of information describing the content of the images (evidence). Information about the content is acquired by utilizing the MPEG-7 descriptors and MD5 hashes. Calculating the descriptors is a time-consuming process. On the other hand, software used in investigations should work automatically. It should also store descriptors and hashes in a local database. As mentioned in [2], in many instances police forces are able to index images simultaneously. Each of the application's instances needs to communicate with a global database and commit new records. These considerations were the main arguments for the creation of the INACT INDEXER [2], [3].

4. INACT Searcher

The purpose of the INACT research system is to make police work more effective. The INACT INSEARCHER is designed and implemented to improve the quality of the police force operations. Special optimization approaches were taken in order to achieve this goal, see [2]. The search process may be performed on a computer belonging to the suspect. The hardware and operating system may be utilized during this procedure therefore the INSEARCHER must run on different operating systems.

5. Implementation

Both the INACT INDEXER and INSEARCHER are written in C programming language, while the Hash Database is based on MySQL. Programming libraries used in the application were carefully chosen to have versions in multiple operating systems. OpenCV [4] was used for image manipulation, QT [5] for GUI and a proprietary MPEG-7 library for similarity measurement [6]. Both applications have Windows and Unix (Ubuntu) versions. For the INACT INDEXER it is expected to be used at the Police premises so it is expected that Windows version will be mostly used. For the INSEARCHER native version for most common operating systems have to be available. For the moment of writing of this paper both Ubuntu and Windows versions of the INSEARCHER are integrated with a DEFT forensic operating system, which is operated from a typical pen drive [7]. GUI of the INACT INDEXER and INACT INSEARCHER is presented in Fig. 2 and Fig. 3 respectively. Both applications are available in English, German, Czech and Polish languages.

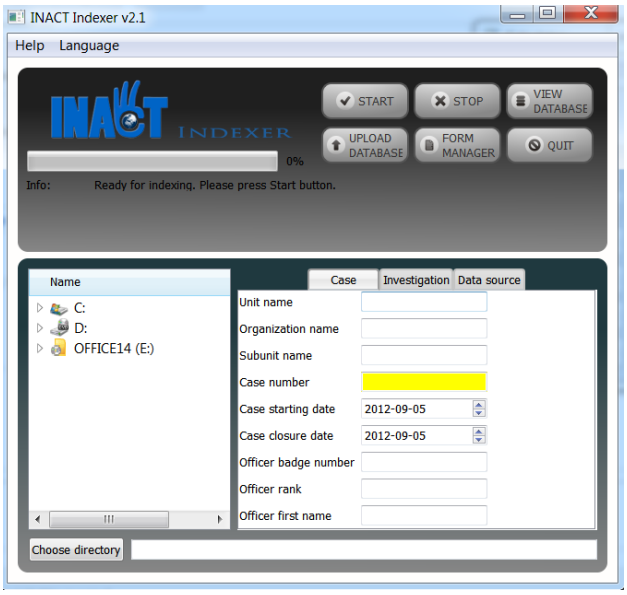

Fig. 2: INACT INDEXER GUI (Windows version).

6. Practical Experiments

Overall performance of the INSEARCHER application was done in [2]. An experiment on a sample test set

including 7151 images of a total size of 11227 MB was conducted.

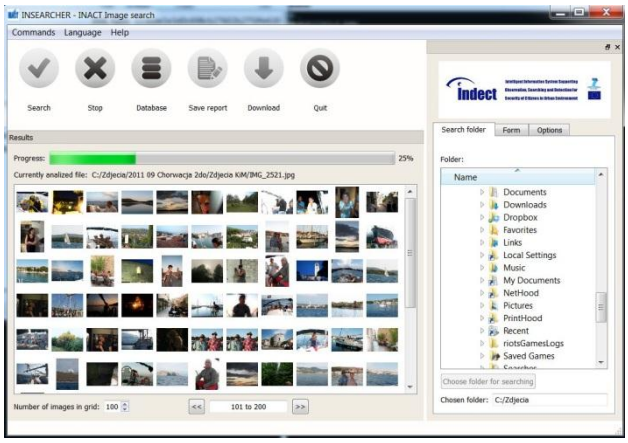

Fig. 3: INACT INSEARCHER GUI (Windows version).

Since the application is primarily being developed for police forces, there is a need to perform a set of experiments which are based on real data. Real data should contain the CP images. Therefore the cooperation with Czech Police, Department of Internet Crime was established. All experiments were performed on the side of Czech Police in line with current law regulations.

In live forensic cases the time to first hit is important, as then further legal actions are justified, e.g. arrest of the suspect and hardware. The "first hit" is the time when the first image containing CP is found. Of course, the Police may continue the searching process in order to find as many CP images as possible, but the first hit really matters.

6.1. General Conditions

For performance evaluation of INSEARCHER tool, it is necessary to know the behaviour of the application. If the search is performed on real data (containing CP images), the first hit have to be evaluated. The suspect images that contain CP can be processed only by police forces therefore, the function for logging first hit times was implemented to the INSEARCHER. Exported log file is in textual form and consists of following parameters:

- FoundFileName,
- SimilarityValue,
- TimeToFind.

FoundFileName is the image ID in the database to which some similarity was found, see [2]. SimilarityValue is an integer and describes how similar is the searched image to that in the database. The actual distance is not displayed in the application, as it is non-informative for the user at all. It is impossible to set up a distinct border-value for the distance metric that would distinguish "similar" from "not similar". This function was implemented only for research purposes. TimeToFind is information how much time did it take to find a given file.

6.2. Results

For practical experiments a few interest images were indexed by INACT INDEXER. Their descriptors are saved to the database according to the rules as described in [2]. The experiment consisted of 10 sets, each set consisted of over 400 images in very high resolution. These sets were subsequently searched by INACT SEARCHER and the log file was stored. Of course, the INACT SEARCHER is fully automatic and therefore all images in the set were processed.

First, identical images are searched. Identical images are identified by comparison of MD5 hashes. If any identical image is found, SimilarityValue is set to zero. Then, it is very easy to extract these records. Nevertheless, the main aim is to determine the "first hit" time in each test set. Lower SimilarityValue represents larger similarity nevertheless it was necessary to extract the records where the real similarity was found. This was done manually. Tab. 1 shows the "first hit" time in each experiment ID. In the first column, the first hit is based on lowest SimilarityValue parameter. In other words, best similarity determined by INSEARCHER. Second column is based on real similarity. This was done manually.

Tab.1: Times of first hit.

Experiment ID	Time to First Hit [ms]	
	Based on SimilarityValue	Based on real similarity
1	50117	16773
2	80094	73295
3	26755	26755
4	76491	81129
5	72755	72755
6	95522	24170
7	149410	179077
8	87501	87501
9	90157	90157
10	60531	60531

If the values in both columns are the same, it means that INSEARCHER found the first similar image correctly. If the values are different in the columns, it means that real similar image had to be recognized manually. The time to first hit can be then smaller or bigger.

7. Conclusion

This paper introduces first real experiments with INACT tool. The experiments were performed on real data, so cooperation with Police was necessary. All real

experiments were performed on the side of Czech Police. The exported log file was analyzed and evaluated. No images were used or available for analysis on the side of Universities. It was proven that same images, which are searched, are recognized with 100 % success. In the case of similarity, another set of experiments are necessary to perform. The experiments should determine on which position the first hit image is mostly located.

Now, it was proved that the similarity was recognized correctly in 5 sets. In 3 cases, the real similarity was recognized earlier than INSEARCHER represented. In 2 cases, the similarity was not correctly recognized by INSEARCHER. These 2 cases had to be recognized manually.

More crucial fact is that all images recognized as the first hit have to be manually checked. Nevertheless, in the case of hundreds of thousand images it could be very beneficial. In next steps, we would like to continue in detailed statistical evaluation of recognized images with the cooperation of Czech Police.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 218086 - project INDECT.

References

[1] Child Pornography: Model Legislation and Global Review. In: *International Centre for Missing and Exploited Children* [online]. 6th ed., 2006. Available at: http://books.google.pl/books?id=8iqqOwAACAAJ.

[2] GREGA, M., D. BRYK and M. NAPORA. INACT- INDECT Advanced Image Cataloguing Tool. *Multimedia Tools and Applications*. 2012. ISSN 1380-7501. DOI: 10.1007/s11042-012-1164-3.

[3] GREGA, M., D. BRYK, M. NAPORA and M. GUSTA. INACT-INDECT advanced image cataloguing tool. *Communications in Computer and Information Science*. 2011, vol. 149. pp. 28-36. ISSN 1865-0929. DOI: 10.1007/978-3-642-21512-4_4.

[4] BRADSKI, Gary. The OpenCV Library. In: *Dr.Dobb's: The world of Software Development* [online]. 2000. Available at: http://www.drdobbs.com/open-source/the-opencv-library/184404319.

[5] JASMIN BLANCHETTE, Mark SUMMERFI. *C++ GUI Programming with Qt 4*. 2nd ed. Westford: Prentice Hall Open Source Software Development Series, 2008. ISBN 978-0132354165.

[6] FRACZEK, R., M. GREGA, N. LIEBAU, M. LESZCZUK, A. LUEDTKE, L. JANOWSKI and Z. PAPIR. Ground-Truth-Less Comparison of Selected Content-Based Image Retrieval Measures. In: *User Centric Media First International Conference (UCMedia 2009)*. Berlin, Heidelberg: Springer-Verlag, 2010, vol. 40, pp. 101-108. ISSN1867-8211. ISBN 978-3-642-12630-7. DOI: 10.1007/978-3-642-12630-7_12.

[7] DEFT Linux: Computer Forensics live cd. In: *DEFT Linux - Computer Forensics live cd* [online]. 2012. Available at: http://www.deftlinux.net.

About Authors

Libor MICHALEK was born in 1979 in Trinec, Czech Republic. He received his M.Sc. from Electronics and Telecommunications in 2004 from the Department of Telecommunications, VSB-Technical University of Ostrava. In 2008 he received Ph.D. from Telecommunications. His professional interest involves optimizing of wireless data systems and modeling and simulation of telecommunication networks.

Michal GREGA was born in 1982 in Krakow, Poland. He received his M.SC. Eng. from the Department of Telecommunications, AGH University of Science and Technology in 2006. In 2011 he received Ph.D. in Telecommunications at AGH. In 2012 he received an IPMA-D Project Management certificate from Krakow University of Economics. His professional interest involves multimedia analysis, understanding and retrieval, quality of multimedia services and project management.

Damian BRYK started his university education at the University of Science and Technology in Krakow, Poland in 2007. In 2011 he presented the Bachelor Thesis "Hearing diagnosis device: audiometer". In 2011 he stated Master studies in Electronics. In 2009 he joined to the INDECT project at the Department of Telecommunications on the Univeristy of Science and Technology. His interest includes programming, embedded systems development and design of electronic devices.

Bartlomiej GRABOWSKI was born in 1989 in Lublin. He received his Eng. from the Department of Applied Computer Science, AGH University of Science and Technology, Krakow, where he now attends a Master course. His professional interests include QT and C++ development.